UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/686,964	10/15/2003	Srinivasan Balasubramanian	ARC920030083US1	8831

28342 7590 06/19/2007
SAMUEL A. KASSATLY LAW OFFICE
20690 VIEW OAKS WAY
SAN JOSE, CA 95120

EXAMINER	
DWIVEDI, MAHESH H	

ART UNIT	PAPER NUMBER
2168	

MAIL DATE	DELIVERY MODE
06/19/2007	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary

Application No.	Applicant(s)	
10/686,964	BALASUBRAMANIAN ET AL.	
Examiner	Art Unit	
Mahesh H. Dwivedi	2168	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133).

Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on 22 March 2007.

2a)☒ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) 1-20 is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) 1-20 is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on 15 October 2003 is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☐ Notice of References Cited (PTO-892)

2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)

3) ☐ Information Disclosure Statement(s) (PTO/SB/08)
Paper No(s)/Mail Date _____.

4) ☐ Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____ .

5) ☐ Notice of Informal Patent Application

6) ☐ Other: _____.

DETAILED ACTION

Remarks

1. Receipt of Applicant's Amendment, filed on 03/22/2007, is acknowledged. The amendment includes the amending of claims 1, 11-14, and 18-20.

Claim Rejections - 35 USC § 112

2. The rejections raised in the office action mailed on 12/27/2006 have been overcome by applicant's amendments received on 03/22/2007.

3. The following is a quotation of the second paragraph of 35 U.S.C. 112:

> The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

4. Claims 1, 14, and 18 are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention. The examiner specifically points to the limitations "identifying a resource locator string associated with the fetched document" and "placing the resource locator string for the fetched document in a blacklist in order to prevent future crawling of the fetched document" as being indefinite. Specifically, the examiner wishes to state that it is unclear as to whether the applicant is referring to the fetched documents which are crawled or the disallowed fetched documents. The examiner suggests that applicant amends the aforementioned limitations by replacing "fetch documents" with "disallowed fetched documents" to potentially overcome the rejections.

Claims 2-13, 15-17, and 19-20 are rejected for incorporating the deficiencies of independent claims 1, 14, and 18.

Claim Rejections - 35 USC § 101

5. The rejections raised in the office action mailed on 12/27/2006 have been overcome by applicant's amendments received on 03/22/2007.

Claim Objections

6. Claim 14 is objected to because of the following informalities: The phrase "wherein the fourth set of instructions codes **further determines**" should be changed to

"wherein the fourth set of instructions codes **further determine**". Appropriate correction is required.

Claims 15-17 are objected to for incorporating the deficiencies of independent claim 14.

Claim Rejections - 35 USC § 103

7. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

8. This application currently names joint inventors. In considering patentability of the claims under 35 U.S.C. 103(a), the examiner presumes that the subject matter of the various claims was commonly owned at the time any inventions covered therein were made absent any evidence to the contrary. Applicant is advised of the obligation under 37 CFR 1.56 to point out the inventor and invention dates of each claim that was not commonly owned at the time a later invention was made in order for the examiner to consider the applicability of 35 U.S.C. 103(c) and potential 35 U.S.C. 102(e), (f) or (g) prior art under 35 U.S.C. 103(a).

9. Claims 1-20 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Chakrabarti et al.** (U.S. Patent 6,418,433) in view of **Liang** (U.S. PGPUB 2001/0044818).

10. Regarding claim 1, **Chakrabarti** teaches a method comprising:

A) selectively prioritizing the documents to crawl based on a set of rules (Column 8, lines 2-30);

B) fetching prioritized documents from the network (Column 5, lines 40-46);

C) for each fetched document, determining whether the fetched document is relevant to any of the multiple focus topics (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 18-43);

D) crawling the fetched document that matches any of the multiple focus topics (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 35-43);

E) further crawling out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 35-43);

F) determined whether the fetched document should be disallowed (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 18-34); and

G) upon determination that the fetched document should be disallowed, selectively disallowing the fetched document (Column 10, lines 18-34);

H) identifying a resource locator string associated with the fetched document (Column 6, lines 66-67-Column 7, lines 1-2);

The examiner notes that **Chakrabarti** teaches **"selectively prioritizing the documents to crawl based on a set of rules"** as "The priority and relevance fields permit two types of crawl policies, i.e., the above-mentioned "soft" and "hard" crawl policies. For the "hard" crawl policy, the classifier 28 is invoked as described above on a Web page, and when it returns the best matching category path, the out-links of the page are entered into the crawl database 30 if and only if some node on the best matching category is marked as "good". FIG. 5 shows the details of such a "hard" crawl policy. As recognized herein, however, such a policy can lead to crawl stagnation, preferred solutions to which are addressed in FIGS. 5 and 6. Alternatively, a "soft" policy can be implemented in which all out-links are entered into the crawl database 30, but their crawl priority is based on the relevance of the current page. A batch of unvisited pages (typically, a few dozen per thread) are selected in lexicographic order of (Num_Tries, relevance desc, priority asc, bytehash), where "asc" means ascending "desc" means descending, and bytehash is a random number to resolve ties without loading any particular server. Each URL from the group is downloaded and classified, which generally leads to a revision of the relevance score. The revised relevance score is also written into the new records created for unvisited out-links" (Column 8, lines 8-30). The examiner further notes that **Chakrabarti** teaches **"fetching prioritized**

documents from the network" as "the Web page table 32 includes a priority field 42 that represents how often the Web page is to be revisited by the crawler 14" (Column 5, lines 41-42). The examiner further notes that **Chakrabarti** teaches **"for each fetched document, determining whether the fetched document is relevant to any of the multiple focus topics"** as "The topic analyzer 28 compares the content of a Web page with a predefined topic or topics and generates a response representative of how relevant the Web page is" (Column 4, lines 61-65), "When the process determines that the page under test is not relevant to the predefined topic" (Column 10, lines 18-19), and "If the page under test is determined to be relevant to the topic" (Column 10, lines 35-36). The examiner further notes that **Chakrabarti** teaches **"crawling the fetched document that matches any of the multiple focus topics"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-39). The examiner further notes that **Chakrabarti** teaches **"further crawling out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-39). The examiner further notes that **Chakrabarti** teaches **"determined whether the fetched document should be disallowed"** as "The topic analyzer 28 compares the content of a Web page with a predefined topic or topics and generates a response representative of how relevant the Web page is" (Column 4, lines 61-65) and "When the process determines that the page under test is not relevant to the predefined topic, the process moves to block 108 to update the Web page table 32 entries for the page under test (if the page is an old page), and then to return to block 86. It is to be understood that only the page under test is recorded at block 108, and that the outlinks of the page under test are not entered into the link table 34. Also, if the page under test is a new but irrelevant page, it is not added to the page table 32 at block 108. Thus, from one aspect, the page under test is pruned at block 108, in that its outlinks are not stored by the system 10 and the page itself is not stored if the page is a new but irrelevant page" (Column 10, lines 18-29). The examiner further notes that **Chakrabarti** teaches **"<u>upon determination that the fetched document should be disallowed, selectively disallowing the fetched document</u>"** as "When the process determines that the page under test is not relevant to the predefined topic, the process moves to block 108 to update the Web page table 32 entries for the page under test (if the page is an old page), and then to return to block 86. It is to be understood that only the page under test is recorded at block 108, and that the outlinks of the page under test are not entered into the link table 34. Also, if the page under test is a new but irrelevant page, it is not added to the page table 32 at block 108. Thus, from one aspect, the page under test is pruned at block 108, in that its outlinks are not stored by the system 10 and the page itself is not stored if the page is a new but irrelevant page" (Column 10, lines 18-29). The examiner further notes that **Chakrabarti** teaches **"<u>identifying a resource locator string associated with the fetched document</u>"** as "From block 108 the process moves to block 110 to tokenize the URL to identify its textual content and out-links. Proceeding to block 112, the relevance and priority of the document are determined as follows" (Column 6, lines 66-67-Column 7, lines 1-2).

Chakrabarti does not explicitly teach:

I) <u>placing the resource locator string for the fetched document in a blacklist in order to prevent future crawling of the fetched document</u>.

Liang, however, teaches **"wherein the miner comprises an unfocus miner that places the resulting uniform resource locator strings that match an unfocus topic in a blacklist, so that the uniform resource locator strings will not be crawled again"** as "web spider 26 determines whether the retrieved web content contains pornographic material. If it does, then in step 908, web spider 26 adds the URL to list 28" (Paragraph 63).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited references because teaching **Liang's** would have allowed **Chakrabarti's** to provide a method to allow for web crawlers and spiders to dynamically restrict unwanted and unacceptable material, as noted by **Liang** (Paragraph 3).

Regarding claim 2, **Chakrabarti** teaches a method comprising:

A) seeding a plurality of seed uniform resource locator strings to start the collaborative focused crawling of the documents (Column 5, lines 61-67-Column 6, lines 1-15).

The examiner notes that **Chakrabarti** teaches **"seeding a plurality of seed uniform resource locator strings to start the collaborative focused crawling of the documents"** as "It is to be understood that information pertaining to a "seed" set of Web pages is initially stored in the Web page table 32. The seed set can be gathered from, e.g., the temporary Internet file directories of the employees of a company or from some other group that can be expected to have shared interests...Thus, the seed set does not define a comprehensive, universal set of all topics on the Web, but rather a relatively narrow topic or range of topics that are of interest to the particular source" (Column 5, lines 61-67-Column 6, lines 1-4).

Regarding claim 3, **Chakrabarti** teaches a method comprising:

A) crawling the seed uniform resource locator strings (Column 6, lines 61-67-Column 7, lines 1-2, Column 10, lines 44-64).

The examiner notes that **Chakrabarti** teaches **"crawling the seed uniform resource locator strings"** as "starting with the seed set the URL of each page is selected" (Column 6, lines 61-62) and "the current page is classified to its topics, using the topic analyzer 28 (FIG. 1), and then the page is evaluated for relevancy to the predefined topic at the decision diamond 116...when the page is a "good" page the logic expands the outlinks of the page" (Column 10, lines 45-51).

Regarding claim 4, **Chakrabarti** teaches a method comprising:

A) writing a plurality of resulting uniform resource locator strings obtained by crawling the seed uniform resource locator strings (Column 10, lines 35-43, 51-64).

The examiner notes that **Chakrabarti** teaches **"writing a plurality of resulting uniform resource locator strings obtained by crawling the seed uniform resource locator strings"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-38).

Regarding claim 5, **Chakrabarti** teaches a method comprising:

A) a foreman function for reading a plurality of contents of the resulting uniform resource locator strings (Column 10, lines 4-10, 51-64)

The examiner notes that **Chakrabarti** teaches **"a foreman function for reading a plurality of contents of the resulting uniform resource locator strings"** as "If the checksum comparison at decision diamond 100 indicates that new data is begin considered, however, the logic proceeds to block 102 to tokenize the Web page" (Column 10, lines 4-6).

Regarding claim 6, **Chakrabarti** teaches a method comprising:

A) the foreman function passing the contents of the resulting uniform resource locator strings to a miner (Column 10, lines 10-17, 51-64).

The examiner notes that **Chakrabarti** teaches **"a foreman function for reading a plurality of contents of the resulting uniform resource locator strings"** as "Then , the page is classified at block 104 using the topic analyzer or classifier 28" (Column 10, lines 10-11).

Regarding claim 7, **Chakrabarti** teaches a method comprising:

A) the miner instructing a fetcher to crawl a plurality of out-links on a document of the resulting resource locator string when the contents of the resulting resource locator string match a focus topic of the miner (Column 10, lines 35-43, 51-64).

The examiner notes that **Chakrabarti** teaches **"the miner instructing a fetcher to crawl a plurality of out-links on a document of the resulting resource locator string when the contents of the resulting resource locator string match a focus topic of the miner"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-38).

Regarding claim 8, **Chakrabarti** teaches a method comprising:

A) the miner ignoring resulting resource locator string when the contents of the resulting resource locator string do not match the focus of the miner (Column 10, lines 18-34).

The examiner notes that **Chakrabarti** teaches **"the miner instructing a fetcher to crawl a plurality of out-links on a document of the resulting resource locator string when the contents of the resulting resource locator string match a focus topic of the miner"** as "When the process determines that the page under test is not relevant to the predefined topic, the process moves to block 108 to update the Web page table 32...the outlinks of the page under test are not entered into the link table" (Column 10, lines 18-24).

Regarding claim 9, **Chakrabarti** teaches a method comprising:

A) the miner managing a plurality of focus topics (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65).

The examiner notes that **Chakrabarti** teaches **"the miner managing a plurality of focus topics"** as "The topic analyzer 28 compares the content of a Web page with a predefined topic or topics and generates a response representative of how relevant the Web page is" (Column 4, lines 61-65).

Regarding claim 14, **Chakrabarti** teaches a computer program product comprising:

A) a first set of instruction codes for selectively prioritizing the documents to crawl based on a set of rules (Column 8, lines 2-30);

B) a second set of instruction codes for fetching prioritized documents from the network (Column 5, lines 40-46);

C) for each fetched document, a third set of instruction codes determines whether the fetched document is relevant to any of the multiple focus topics (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 18-43);

D) a fourth set of instruction codes for crawling the fetched document that matches any of the multiple focus topics (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 35-43);

E) wherein the fourth set of instruction codes further crawls out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 35-43);

F) wherein the fourth set of instruction codes further determines whether the fetched document should be disallowed (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 18-34); and

G) upon determination that the fetched document should be disallowed, selectively disallowing the fetched document (Column 10, lines 18-34);

H) identifying a resource locator string associated with the fetched document (Column 6, lines 66-67-Column 7, lines 1-2);

The examiner notes that **Chakrabarti** teaches **"a first set of instruction codes for selectively prioritizing the documents to crawl based on a set of rules"** as "The priority and relevance fields permit two types of crawl policies, i.e., the above-mentioned "soft" and "hard" crawl policies (Column 8, lines 8-11). The examiner further notes that **Chakrabarti** teaches **"a second set of instruction codes for fetching prioritized documents from the network"** as "the Web page table 32 includes a priority field 42 that represents how often the Web page is to be revisited by the crawler 14" (Column 5, lines 41-42). The examiner further notes that **Chakrabarti** teaches **"for each fetched document, a third set of instruction codes determines whether the fetched document is relevant to any of the multiple focus topics"** as "The topic analyzer 28 compares the content of a Web page with a predefined topic or topics and generates a response representative of how relevant the Web page is" (Column 4, lines 61-65), "When the process determines that the page under test is not relevant to the predefined topic" (Column 10, lines 18-19), and "If the page under test is determined to be relevant to the topic" (Column 10, lines 35-36). The examiner further notes that **Chakrabarti** teaches **"a fourth set of instruction codes for crawling the fetched document that**

matches any of the multiple focus topics" as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-39). The examiner further notes that **Chakrabarti** teaches **"wherein the fourth set of instruction codes further crawls out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-39). The examiner further notes that **Chakrabarti** teaches **"<u>wherein the fourth set of instruction codes further determines whether the fetched document should be disallowed</u>"** as "The topic analyzer 28 compares the content of a Web page with a predefined topic or topics and generates a response representative of how relevant the Web page is" (Column 4, lines 61-65) and "When the process determines that the page under test is not relevant to the predefined topic, the process moves to block 108 to update the Web page table 32 entries for the page under test (if the page is an old page), and then to return to block 86. It is to be understood that only the page under test is recorded at block 108, and that the outlinks of the page under test are not entered into the link table 34. Also, if the page under test is a new but irrelevant page, it is not added to the page table 32 at block 108. Thus, from one aspect, the page under test is pruned at block 108, in that its outlinks are not stored by the system 10 and the page itself is not stored if the page is a new but irrelevant page" (Column 10, lines 18-29). The examiner further notes that **Chakrabarti** teaches **"<u>upon determination that the fetched document should be disallowed, selectively disallowing the fetched document</u>"** as "When the process determines that the page under test is not relevant to the predefined topic, the process moves to block 108 to update the Web page table 32 entries for the page under test (if the page is an old page), and then to return to block 86. It is to be understood that only the page under test is recorded at block 108, and that the outlinks of the page under test are not entered into the link table 34. Also, if the page under test is a new but irrelevant page, it is not added to the page table 32 at block 108. Thus, from one aspect, the page under test is pruned at block 108, in that its outlinks are not stored by the system 10 and the page itself is not stored if the page is a new but irrelevant page" (Column 10, lines 18-29). The examiner further notes that **Chakrabarti** teaches **"identifying a resource locator string associated with the fetched document"** as "From block 108 the process moves to block 110 to tokenize the URL to identify its textual content and out-links. Proceeding to block 112, the relevance and priority of the document are determined as follows" (Column 6, lines 66-67-Column 7, lines 1-2).

Chakrabarti does not explicitly teach:

I) placing the resource locator string for the fetched document in a blacklist in order to prevent future crawling of the fetched document.

Liang, however, teaches **"wherein the miner comprises an unfocus miner that places the resulting uniform resource locator strings that match an unfocus topic in a blacklist, so that the uniform resource locator strings will not be crawled again"** as "web spider 26 determines whether the retrieved web content contains pornographic material. If it does, then in step 908, web spider 26 adds the URL to list 28" (Paragraph 63).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited references because teaching **Liang's** would have allowed **Chakrabarti's** to provide a method to allow for web crawlers and spiders to dynamically restrict unwanted and unacceptable material, as noted by **Liang** (Paragraph 3).

Regarding claim 15, **Chakrabarti** teaches a computer program product comprising:

A) a fifth set of instruction codes for seeding a plurality of seed uniform resource locator strings to start the collaborative focused crawling of the documents (Column 5, lines 61-67-Column 6, lines 1-15).

The examiner notes that **Chakrabarti** teaches **"a fifth set of instruction codes for seeding a plurality of seed uniform resource locator strings to start the collaborative focused crawling of the documents"** as "It is to be understood that information pertaining to a "seed" set of Web pages is initially stored in the Web page table 32. The seed set can be gathered from, e.g., the temporary Internet file directories of the employees of a company or from some other group that can be expected to have shared interests...Thus, the seed set does not define a comprehensive, universal set of all topics on the Web, but rather a relatively narrow topic or range of topics that are of interest to the particular source" (Column 5, lines 61-67-Column 6, lines 1-4).

Regarding claim 16, **Chakrabarti** teaches a computer program product comprising:

A) wherein the fourth set of instruction codes further crawls the seed uniform resource locator strings (Column 6, lines 61-67-Column 7, lines 1-2, Column 10, lines 44-64).

The examiner notes that **Chakrabarti** teaches **"wherein the fourth set of instruction codes further crawls the seed uniform resource locator strings"** as "starting with the seed set the URL of each page is selected" (Column 6, lines 61-62) and "the current page is classified to its topics, using the topic analyzer 28 (FIG. 1), and then the page is evaluated for relevancy to the predefined topic at the decision diamond 116...when the page is a "good" page the logic expands the outlinks of the page" (Column 10, lines 45-51).

Regarding claim 17, **Chakrabarti** teaches a computer program product comprising:

A) a sixth set of instruction codes for writing a plurality of resulting uniform resource locator strings obtained by crawling the seed uniform resource locator strings (Column 10, lines 35-43, 51-64).

The examiner notes that **Chakrabarti** teaches **"a sixth set of instruction codes for writing a plurality of resulting uniform resource locator strings obtained by crawling the seed uniform resource locator strings"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-38).

Regarding claim 18, **Chakrabarti** teaches a system comprising:

A) an evaluator that selectively prioritizes the documents to crawl based on a set of rules (Column 8, lines 2-30);

B) a fetcher that fetches prioritized documents from the network (Column 5, lines 40-46);

C) for each fetched document, a focus engine determines whether the fetched document is relevant to any of the multiple focus topics (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 18-43);

D) a crawler for crawling the fetched document that matches any of the multiple focus topics (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 35-43);

E) wherein the crawler further crawls out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 35-43);

F) wherein the crawler further determines whether the fetched document should be disallowed (Column 2, lines 56-60, Column 3, lines 51-55, Column 4, lines 61-65, Column 10, lines 18-34); and

G) upon determination that the fetched document should be disallowed, selectively disallowing the fetched document (Column 10, lines 18-34);

H) identifying a resource locator string associated with the fetched document (Column 6, lines 66-67-Column 7, lines 1-2);

The examiner notes that **Chakrabarti** teaches **"an evaluator that selectively prioritizes the documents to crawl based on a set of rules"** as "the Web page table 32 includes a priority field 42 that represents how often the Web page is to be revisited by the crawler 14" (Column 5, lines 41-42). The examiner further notes that **Chakrabarti** teaches **"for each fetched document, a focus engine determines**

whether the fetched document is relevant to any of the multiple focus topics" as "The topic analyzer 28 compares the content of a Web page with a predefined topic or topics and generates a response representative of how relevant the Web page is" (Column 4, lines 61-65), "When the process determines that the page under test is not relevant to the predefined topic" (Column 10, lines 18-19), and "If the page under test is determined to be relevant to the topic" (Column 10, lines 35-36). The examiner further notes that **Chakrabarti** teaches **"a crawler for crawling the fetched document that matches any of the multiple focus topics"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-39). The examiner further notes that **Chakrabarti** teaches **"wherein the crawler further crawls out-links on the fetched document based on an assumption that if the fetched document is of interest, the out-links are also of interest"** as "If the page under test is determined to be relevant to the topic, however, the process moves to block 110, wherein entries are generated for the link table 34 for all outlinks of the page" (Column 10, lines 35-39). The examiner further notes that **Chakrabarti** teaches **"**<u>wherein the crawler further determines whether the fetched document should be disallowed</u>**"** as "The topic analyzer 28 compares the content of a Web page with a predefined topic or topics and generates a response representative of how relevant the Web page is" (Column 4, lines 61-65) and "When the process determines that the page under test is not relevant to the predefined topic, the process moves to block 108 to update the Web page table 32 entries for the page under test (if the page is an old page), and then to return to block 86. It is to be understood that only the page under test is recorded at block 108, and that the outlinks of the page under test are not entered into the link table 34. Also, if the page under test is a new but irrelevant page, it is not added to the page table 32 at block 108. Thus, from one aspect, the page under test is pruned at block 108, in that its outlinks are not stored by the system 10 and the page itself is not stored if the page is a new but irrelevant page" (Column 10, lines 18-29). The examiner further notes that **Chakrabarti** teaches **"<u>upon determination that the fetched document should be disallowed, selectively disallowing the fetched</u>**

document" as "When the process determines that the page under test is not relevant to the predefined topic, the process moves to block 108 to update the Web page table 32 entries for the page under test (if the page is an old page), and then to return to block 86. It is to be understood that only the page under test is recorded at block 108, and that the outlinks of the page under test are not entered into the link table 34. Also, if the page under test is a new but irrelevant page, it is not added to the page table 32 at block 108. Thus, from one aspect, the page under test is pruned at block 108, in that its outlinks are not stored by the system 10 and the page itself is not stored if the page is a new but irrelevant page" (Column 10, lines 18-29). The examiner further notes that **Chakrabarti** teaches "**identifying a resource locator string associated with the fetched document**" as "From block 108 the process moves to block 110 to tokenize the URL to identify its textual content and out-links. Proceeding to block 112, the relevance and priority of the document are determined as follows" (Column 6, lines 66-67-Column 7, lines 1-2).

Chakrabarti does not explicitly teach:

I) placing the resource locator string for the fetched document in a blacklist in order to prevent future crawling of the fetched document.

Liang, however, teaches **"wherein the miner comprises an unfocus miner that places the resulting uniform resource locator strings that match an unfocus topic in a blacklist, so that the uniform resource locator strings will not be crawled again"** as "web spider 26 determines whether the retrieved web content contains pornographic material. If it does, then in step 908, web spider 26 adds the URL to list 28" (Paragraph 63).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited references because teaching **Liang's** would have allowed **Chakrabarti's** to provide a method to allow for web crawlers and spiders to dynamically restrict unwanted and unacceptable material, as noted by **Liang** (Paragraph 3).

Regarding claim 19, **Chakrabarti** teaches a system comprising:

A) a plurality of seed uniform resource locator strings that are used to initiate the collaborative focused crawling of the documents (Column 5, lines 61-67-Column 6, lines 1-15).

The examiner notes that **Chakrabarti** teaches **"a plurality of seed uniform resource locator strings that are used to initiate the collaborative focused crawling of the documents"** as "It is to be understood that information pertaining to a "seed" set of Web pages is initially stored in the Web page table 32. The seed set can be gathered from, e.g., the temporary Internet file directories of the employees of a company or from some other group that can be expected to have shared interests...Thus, the seed set does not define a comprehensive, universal set of all topics on the Web, but rather a relatively narrow topic or range of topics that are of interest to the particular source" (Column 5, lines 61-67-Column 6, lines 1-4).

Regarding claim 20, **Chakrabarti** teaches a system product comprising:

A) wherein the crawler further crawls the seed uniform resource locator strings (Column 6, lines 61-67-Column 7, lines 1-2, Column 10, lines 44-64).

The examiner notes that **Chakrabarti** teaches "wherein the crawler further crawls the seed uniform resource locator strings" as "starting with the seed set the URL of each page is selected" (Column 6, lines 61-62) and "the current page is classified to its topics, using the topic analyzer 28 (FIG. 1), and then the page is evaluated for relevancy to the predefined topic at the decision diamond 116...when the page is a "good" page the logic expands the outlinks of the page" (Column 10, lines 45-51).

Claim Rejections - 35 USC § 103

11. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

This application currently names joint inventors. In considering patentability of the claims under 35 U.S.C. 103(a), the examiner presumes that the subject matter of the various claims was commonly owned at the time any inventions covered therein were made absent any evidence to the contrary. Applicant is advised of the obligation under 37 CFR 1.56 to point out the inventor and invention dates of each claim that was not commonly owned at the time a later invention was made in order for the examiner to consider the applicability of 35 U.S.C. 103(c) and potential 35 U.S.C. 102(e), (f) or (g) prior art under 35 U.S.C. 103(a).

12. Claims 10-13 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Chakrabarti et al.** (U.S. Patent 6,418,433) in view of **Liang** (U.S. PGPUB 2001/0044818) as applied to claims 1-9, and 14-20, and in view of **Heydon et al.** (Article entitled "Mercator: A Scalable, Extensible Web Crawler", dated 06/26/1999).

13. Regarding claim 10, **Chakrabarti** and **Liang** do not explicitly teach a method comprising:

A) the miner allowing a crawling of the resulting resource locator string when the resulting resource locator string matches a plurality of web space rules.

Heydon, however, teaches **"the miner allowing a crawling of the resulting resource locator string when the resulting resource locator string matches a plurality of web space rules"** as "The URL filtering mechanism provides a customizable way to control the set of URLs that are downloaded...The URL filter class has a single crawl method that takes a URL and returns a Boolean value indicating whether or not to crawl that URL" (Page 6, Section: 3.6: URL Filters).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited references because teaching **Heydon's** would have allowed **Chakrabarti's** and **Liang's** to provide a scalable and customizable web crawler to fit a specific user's needs, as noted by **Heydon** (Page 2, Section: 2: Related Work).

Regarding claim 11, **Chakrabarti** and **Liang** do not explicitly teach a method comprising:

A) wherein the web space rules comprise domain rules, IP address rules, and prefix rules.

Heydon, however, teaches **"wherein the web space rules comprise domain rules, IP address rules, and prefix rules"** as "Mercator includes a collection of different URL filter subclasses that provide facilities for restricting URLs by domain, prefix, or protocol type" (Page 6, Section: 3.6: URL Filters).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited references because teaching **Heydon's** would have allowed **Chakrabarti's** and **Liang's** to provide a scalable and customizable web crawler to fit a specific user's needs, as noted by **Heydon** (Page 2, Section: 2: Related Work).

Regarding claim 12, **Chakrabarti** does not explicitly teach a method comprising:
A) the miner disallowing the crawling of the resulting resource locator string when the content of the resulting resource locator string matches a focus topic of the miner.

Liang, however, teaches **"the miner disallowing the crawling of the resulting resource locator string when the content of the resulting resource locator string matches a focus topic of the miner"** as "Web spider 26 is preferably provided with a copy of the lexicon described above so as to permit it to recognize pornographic material" (Paragraph 62) and "if any page in a website is discovered as comprising pornographic material, all pages "below" that page in the sitemap for the website may be blocked (Paragraph 68).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited references because teaching **Liang's** would have allowed **Chakrabarti's** to provide a method to allow for web crawlers and spiders to dynamically restrict unwanted and unacceptable material, as noted by **Liang** (Paragraph 3).

Regarding claim 13, **Chakrabarti** does not explicitly teach a method comprising:

A) wherein the miner comprises an unfocus miner that places the resulting uniform resource locator strings that match an unfocus topic in the blacklist, so that the uniform resource locator strings will not be crawled again.

Liang, however, teaches **"wherein the miner comprises an unfocus miner that places the resulting uniform resource locator strings that match an unfocus topic in the blacklist, so that the uniform resource locator strings will not be crawled again"** as "web spider 26 determines whether the retrieved web content contains pornographic material. If it does, then in step 908, web spider 26 adds the URL to list 28" (Paragraph 63).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited references because teaching **Liang's** would have allowed **Chakrabarti's** to provide a method to allow for web crawlers and spiders to dynamically restrict unwanted and unacceptable material, as noted by **Liang** (Paragraph 3).

Response to Arguments

14. Applicant's arguments filed on 03/22/2007 have been fully considered but they are not persuasive.

Applicant goes on to argue on page 9, that **"Applicants agree with the Examiner that Chakrabarti does not explicitly teach: selectively prioritizing the documents on a set of rules. However, the Examiner notes a contradictory position, namely that Chakrabarti teaches selectively prioritizing the documents to crawl based on a set of rules"**. However, the examiner wishes to state that the aforementioned was a typographical error resultant from a cut and paste operation. Furthermore the examiner wishes to state that the examiner clearly intended to show that **Chakrabarti** taught "selectively prioritizing the documents on a set of rules" as shown with the rejections of independent claims 14 and 18 (each of which do not contain the typographical error). Moreover, the examiner wishes to state that as a result of the amendments made to each of the independent claims, this office action will be made as necessitated by those amendments.

Applicant goes on to argue on page 9, that **"Contrary to Chakrabarti, the instant claim 1 recites "selectively prioritizing the documents to crawl". In other words, the documents are selectively prioritized prior to crawling"**. However, the examiner wishes to point to Column 8 of **Chakrabarti** which states "The priority and relevance fields permit two types of crawl policies, i.e., the above-mentioned "soft" and "hard" crawl policies. For the "hard" crawl policy, the classifier 28 is invoked as described above on a Web page, and when it returns the best matching category path, the out-links of the page are entered into the crawl database 30 if and only if some node on the best matching category is marked as "good". FIG. 5 shows the details of such a "hard" crawl policy. As recognized herein, however, such a policy can lead to crawl stagnation, preferred solutions to which are addressed in FIGS. 5 and 6. Alternatively, a "soft" policy can be implemented in which all out-links are entered into the crawl database 30, but their crawl priority is based on the relevance of the current page. A batch of unvisited pages (typically, a few dozen per thread) are selected in lexicographic order of (Num_Tries, relevance desc, priority asc, bytehash), where "asc" means ascending "desc" means descending, and bytehash is a random number to resolve ties without loading any particular server. Each URL from the group is downloaded and classified, which generally leads to a revision of the relevance score. The revised relevance score is also written into the new records created for unvisited out-links" (Column 8, lines 8-30). The examiner further wishes to state that it is clear that **Chakrabarti's** method has differing levels of priority before crawling a web page (see "hard" and "soft"). The examiner further wishes to state that **Chakrabarti's** method clearly varies the crawling policy (hard or soft) based on the storage space of the system before crawling a new page (hard results in only the best URL's out-links being parsed, whereas soft results in every out-link is parsed).

Applicant goes on to argue on pages 9-10, that **"In addition, the current method prioritizes the documents rather than classifies them as described in Chakrabarti. Classification does not necessarily imply order the documents in a predetermined order"**. However, the examiner wishes to point to Column 8 of **Chakrabarti** which states "The priority and relevance fields permit two types of crawl policies, i.e., the above-mentioned "soft" and "hard" crawl policies. For the "hard" crawl policy, the classifier 28 is invoked as described above on a Web page, and when it returns the best matching category path, the out-links of the page are entered into the crawl database 30 if and only if some node on the best matching category is marked as "good". FIG. 5 shows the details of such a "hard" crawl policy. As recognized herein, however, such a policy can lead to crawl stagnation, preferred solutions to which are addressed in FIGS. 5 and 6. Alternatively, a "soft" policy can be implemented in which all out-links are entered into the crawl database 30, but their crawl priority is based on the relevance of the current page. A batch of unvisited pages (typically, a few dozen per thread) are selected in lexicographic order of (Num_Tries, relevance desc, priority asc, bytehash), where "asc" means ascending "desc" means descending, and bytehash is a random number to resolve ties without loading any particular server. Each URL from the group is downloaded and classified, which generally leads to a revision of the relevance score. The revised relevance score is also written into the new records created for unvisited out-links" (Column 8, lines 8-30). The examiner further wishes to state that it is clear that **Chakrabarti's** method has differing levels of priority before crawling a web page (see "hard" and "soft"). The examiner further wishes to state that **Chakrabarti's** method clearly varies the crawling policy (hard or soft) based on the storage space of the system before crawling a new page (hard results in only the best URL's out-links being parsed, whereas soft results in every out-link is parsed). In response to applicant's argument that the references fail to show certain features of applicant's invention, it is noted that the features upon which applicant relies (i.e., the documents in a predetermined order) are not recited in the rejected claim(s). Although the claims are interpreted in light of the specification, limitations from the specification are not read into the claims. See *In re Van Geuns*, 988 F.2d 1181, 26 USPQ2d 1057 (Fed. Cir. 1993).

Applicant argues on pages 12-13, that **"The google search engine is a...four and eight machines". To the best understanding of the Applicant, Heydon describes downloading web pages and then reads the URLs contained therein. Contrary to Heydon, claim 10 recited the following element: "comprising the miner allowing a crawling of the resulting resource locator string when the**

resulting resource locator string matches a plurality of web space rules". In other terms, the order of operation recited in claim 10 is opposite to that in Heydon. More specifically, according to claim 10, the resulting resource locator string first matches to a plurality of web space rules and then, based on such matching step, allows the crawling of the resulting resource locator string". However, the examiner wishes to state that the cited reference to the google application is that of related work in the **Heydon** reference, and is not remotely related to **Heydon's** method. Moreover, the examiner further wishes to point to page 6 of **Heydon** which states "The URL filtering mechanism provides a customizable way to control the set of URLs that are downloaded. Before adding a URL to the frontier, the worker thread consults the user-supplied URL filter. The URL filter class has a single *crawl* method that takes a URL and returns a boolean value indicating whether or not to crawl that URL. Mercator includes a collection of different URL filter subclasses that provide facilities for restricting URLs by domain, prefix, or protocol type, and for computing the conjunction, disjunction, or negation of other filters. Users may also supply their own custom URL filters, which are dynamically loaded at start-up" (Page 6, Section: 3.6: URL Filters). The examiner further wishes to state that it is clear that **Heydon's** method clearly uses a filter to determine whether or not to crawl a URL based on a received Boolean variable (rules).

Applicant argues on page 14, that **"Applicants respectfully submit that Heydon does not provide any teaching or suggestion to support modifying the Chakrabarti design, as presented earlier in support of the allowance of claim 1"**. In response to applicant's argument that there is no suggestion to combine the references, the examiner recognizes that obviousness can only be established by combining or modifying the teachings of the prior art to produce the claimed invention where there is some teaching, suggestion, or motivation to do so found either in the references themselves or in the knowledge generally available to one of ordinary skill in the art. See *In re Fine*, 837 F.2d 1071, 5 USPQ2d 1596 (Fed. Cir. 1988) and *In re Jones*, 958 F.2d 347, 21 USPQ2d 1941 (Fed. Cir. 1992). In this case, the specified motivation of providing a scalable and customizable web crawler to fit a specific user's needs (Page 2, Section: 2: Related Work), clearly provides **Chakrabarti** with the ability for user customization in URL filtering.

Applicant argues on page 14, that **"the Examiner provided a general reason for the desirability of the combination of Chakrabarti and Heydon, in hindsight, without referring to any sustentative (or significant) teaching or suggestion in Heydon in support of such combination. More specifically, the foregoing reason provided by the examiner is generic, insufficiently specific, and does not provided any ground for the combination of Chakrabarti and Heydon** ".In response to applicant's argument that the examiner's conclusion of obviousness is based upon improper hindsight reasoning, it must be recognized that any judgment on obviousness is in a sense necessarily a reconstruction based upon hindsight reasoning. But so long as it takes into account only knowledge which was within the level of ordinary skill at the time the claimed invention was made, and does not include knowledge gleaned only from the applicant's disclosure, such a reconstruction is proper. See *In re McLaughlin*, 443 F.2d 1392, 170 USPQ 209 (CCPA 1971).

Conclusion

15. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

U.S. Patent 6,199,081 issued to **Meyerzon et al.** on 06 March 2001. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. PGPUB 2004/0049514 issued to **Burkov** on 11 March 2004. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. PGPUB 2002/0194161 issued to **McNamee et al.** on 19 December 2002. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. Patent 6,754,873 issued to **Law et al.** on 22 June 2002. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. Patent 7,080,073 issued to **Jiang et al.** on 18 July 2006. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. PGPUB 2006/0277175 issued to **Jiang et al.** on 07 December 2006. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. Patent 6,993,534 issued to **Denesuk et al.** on 31 January 2006. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. Patent 6,295,559 issued to **Emens et al.** on 25 September 2001. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

U.S. PGPUB 2002/0032869 issued to **Lamberton et al.** on 14 March 2002. The subject matter disclosed therein is pertinent to that of claims 1-20 (e.g., methods to specifically crawl targeted subject matter).

16. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP § 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

Contact Information

17. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Mahesh Dwivedi whose telephone number is (571) 272-2731. The examiner can normally be reached on Monday to Friday 8:20 am – 4:40 pm.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Tim Vo can be reached (571) 272-3642. The fax number for the organization where this application or proceeding is assigned is (571) 273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

Mahesh Dwivedi
Patent Examiner
Art Unit 2168

TIM VO
SUPERVISORY PATENT EXAMINER
TECHNOLOGY CENTER 2100

June 07, 2007